This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(19) Europäisch s Patentamt
European Patent Office
Offic uropéen d s br vets

(11) EP 0 750 293 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
27.12.1996 Bulletin 1996/52

(51) Int Cl.6: G10L 5/06, G10L 7/08,
G10L 9/06

(21) Application number: 96304526.5

(22) Date of filing: 18.06.1996

(84) Designated Contracting States:
DE FR GB IT

(30) Priority: 19.06.1995 JP 151489/95

(71) Applicant: CANON KABUSHIKI KAISHA
Tokyo (JP)

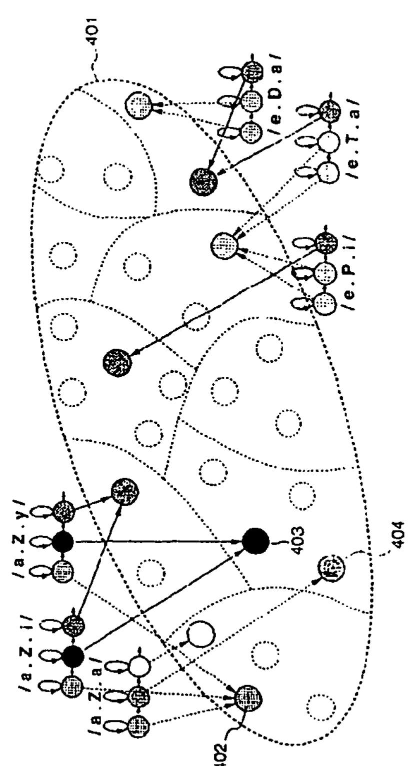
(72) Inventors:
- Komori, Yasuhiro
Ohta-ku, Tokyo (JP)
- Ohora, Yasunori
Ohta-ku, Tokyo (JP)

(74) Representative:
Beresford, Keith Denis Lewis et al
BERESFORD & Co.
2-5 Warwick Court
High Holborn
London WC1R 5DJ (GB)

(54) State transition model design method and voice recognition method and apparatus using same

(57) An object of the invention is to provide a method of generating a state transition model capable of high speed voice recognition and to provide a voice recognition method and apparatus using the state transition model. To this end, a method is provided which generates a state transition model in which a state shared structure of the state transition model is designed, the method including a step of setting the states of a triphone state transition model in an acoustic space as initial clusters, a clustering step of generating a cluster containing the initial clusters by top-down clustering, a step of determining a state shared structure by assigning a short distance cluster among clusters generated by the clustering step, to the state transition model and a step of learning a state shared model by analyzing the states of the triphones in accordance with the determined state shared structure.

FIG. 4



EP 0 750 293 A2

Description

The present invention relates to a design method for a state transition model used, for example, for a voice recognition model. The present invention also relates to a voice recognition method and apparatus using a state transition model designed to recognize voices at high speed.

In the following, a hidden Markov model (HMM) is used as a voice recognition model by way of example.

As the speed of a computer has drastically increased, studies on practical uses and production of voice recognition systems have been made extensively. These systems incorporate HMM which is a statistical model. A triphone HMM has been widely studied because this shows better performance than other HMMs. With this triphone HMM, differences in phone environments such as preceding and succeeding phones are classified finely. This triphone HMM has a number of models so that trainability of data degrades and models of high performance cannot be configured. Furthermore, the computation amount becomes large in proportion with the number of models, posing a critical issue on voice recognition which is required to process in real time.

Several methods for solving these problems have been studied basing upon a concept of "shared structure HMM".

(1) A generalized triphone HMM which shares HMMs themselves having similar acoustic characteristics of the whole phone section (K. F. Lee, H. W. Hon, Large-vocabulary speaker-independent continuous speech recognition using HMM, ICASSP88, pp. 123-126)

(2) A shared-state HMM which shares the states of HMMs having similar acoustic characteristics of the whole phone section (Mei-yuh Hwang, X. D. Huang, Subphonetic modelling with Markov States - SENON, ICASSP92, pp. 133-136, S. J. Young, P. Woodland, The use of state tying in continuous speech recognition, Eurospeech 93, pp. 2203-2206, 1993).

(3) A tied mixture HMM which shares the distributions of HMMs having similar acoustic characteristics of the whole phone section (J. Bellegarda, D. Nahamoo, Tied mixture continuous parameter models for large vocabular isolated speech recognition, ICASSP89, pp. 13-16, D. Paul, The Lncoln robust continuous speech rocognition, ICASSP89, pp. 449-452).

Of these and others, a shared-state HMM using successive state splitting (SSS) proposed by Takami and realizing both the above (1) and (2) is known as a method of generating a shared-state triphone HMM of high precision because a shared state is determined in a top-down manner while considering phone environments (refer to Takami, Sagayama: "Automatic generation of hidden Markov network by SSS", Papers of the Institute of Electronics, Information and Communication Engineers, J76-DII, No. 10, pp. 2155-2164, 1993).

X. D. Huang, S. J. Young, et al. have proposed a method of generating a shared-state triphone HMM through bottom-up merge and obtained good results. Takahashi, et al. have proposed a method of generating an HMM which method synthesizes the above (1) to (3) (refer to Takahashi, Sagayama: "HMM for four hierarchical-level shared structure", Technical Reports of the Institute of Electronics, Information and Communication Engineers, SP94-73, pp. 25-32, 1994-12).

In this invention, all triphones are prepared and the states of these triphones are clustered. In this context, it is analogous to the methods by X. D. Huang and S. J. Young. However, different from clustering through merge considering only local likelihood, top-down clustering considering the whole acoustic space is performed and this clustering is efficient because of consideration of the whole acoustic space.

Although the same top-down scheme as SSS is used, SSS has an inefficient point that an ending state of one triphone is not shared with a starting state of another triphone because of successive state splitting (SSS). Since voices are generally continuously converted, it can be considered relatively natural that a connectable ending state of a triphone and the starting state of the next triphone are to be shared. The method by S. J. Young considers a share of only states within a phone class and cannot share states between phone classes. These disadvantages of SSS have been solved by Takami by incorporating merge into the processes of successive splitting (refer to Takami, "Efficiency improvement of hidden Marcov network by state splitting method", Papers of Lectures of Acoustical Society of Japan, 1-8-4, pp. 7-8, 1994-10). Takahashi and et al. have solved the above disadvantages by incorporating a tied-mixtured HMM. However, the present inventors consider more desirable that the above disadvantages are to be solved from the viewpoint of a state level.

Another disadvantage of SSS is that if an arbitrary speaker HMM is generated by successive state splitting, this splitting becomes dependent upon the arbitrary speaker. It is therefore necessary to use a specified speaker in obtaining a state shared structure. This poses other problems that a large amount of data is required for the specified speaker and that it is necessary to use the state shared structure of the specified speaker for other arbitrary speakers.

The invention has been made under the above circumstances. According to one aspect, the present invention aims to provide a state transition model design method and apparatus capable of recognizing voices at high speed, and a voice recognition method and apparatus sing the state transition model.

BNSDOCID <EP_ 0750293A2 I >

According to another aspect, the present invention aims to provide a state transition model design method and apparatus capable of sharing states between phone lasses or within a phone class, and a voice recognition ethod and apparatus using the state transition model.

According to another aspect, the present invention ims to provide a state transition model design method nd apparatus capable of obtaining a state shared tructure of phones of an arbitrary speaker and capable of efficiently designing a state transition model, and a voice recognition method and apparatus using the state transition model.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a flow chart illustrating processes according to a first embodiment of the invention.

Fig. 2 shows states of an HMM and a state transition model diagram.

Fig. 3 is a flow chart illustrating top-down clustering processes.

Fig. 4 is a diagram illustrating a state shared type HMM.

Fig. 5 is a block diagram illustrating a voice recognition process used by a voice recognition apparatus of the embodiment.

Fig. 6 is a table showing the results of recognition of 100 sentences spoken by 10 arbitrary speakers, the recognition being made by using grammars constituted by 1000 words and the voice recognition apparatus of the embodiment.

Fig. 7 is a flow chart illustrating processes by a second embodiment.

DETAILED DESCRIPTION OF THE EXEMPLARY EMBODIMENTS

Exemplary embodiments of the present invention will be described in detail with reference to the accompanying drawings.

The embodiments (inclusive of flow charts) of the invention are reduced in practice under the control of a CPU in accordance with a control program stored in a ROM or RAM. This control program may be stored in a removable storage medium such as a CD-ROM mounted on a voice recognition apparatus.

A method of designing a state shared structure model for voice recognition according to a first embodiment of the invention will be described.

Fig. 1 is a flow chart illustrating the processes of the first embodiment.

Referring to Fig. 1, reference numeral 101 represents a means (process) for designing initial clusters, reference numeral 102 represents a means (process) for top-down clustering such as general LBG for generating clusters raised to a power of 2, i.e., a means (process) for finely classifying clusters starting from a small number of clusters and sequentially increasing the number of clusters, reference numeral 103 represents a means (process) for determining a common status structure (or state shared structure) of a triphone HMM (modeling considering both preceding and succeeding phones), and reference numeral 104 represents a means (process) for studying (learning) a triphone HMM of the state shared structure.

The details of these means (processes) will be described.

(1) Design of Initial Clusters (101)

(A) All triphone HMMs are learnt using data of an arbitrary speaker.

(a) Phone HMMs of one distribution are learnt at the proper number of states.

(b) A right environment type (right-context) HMM is learnt by using the phone HMMs as initial models.

(c) A both-side environment type (triphone) HMM is learnt by using the right-context HMMs as initial models.

(B) All states of triphone HMMs are used as initial clusters.

Fig. 2 is a diagram illustrating HMM and showing a general state and a state transition model.

In Fig. 2, a state transition probability is indicated by a, an output probability at the corresponding state is indicated by b, a mean value of output probabilities is indicated by μ, and a corresponding dispersion is indicated by σ.

(2) Top-down Clustering by LBG scheme (102)

The top-down clustering is performed by an LBG scheme using a distance scale considering the output probability distribution. Clustering is defined only by the output probability b which is considered to be an important parameter for obtaining a likelihood to HMMs, by neglecting the state transition probability a.

This process is illustrated in the flow chart of Fig 3

3NSDOCID: <EP_ 0750293A2_I_>

At Step S1, 1 is set to m. At Step S2, one class Φm is generated which contains all initial clusters (φi). At Step S3 it is checked if the value m is equal to the total number M (e.g., 600) of clusters. If equal, the process is terminated, whereas if not, the process advances to Step S4.

At Step S4, a new cluster Φm is generated from all the initial clusters (φi) belonging to the old cluster Φm by using the following equations (1) and (2). Specifically, the new cluster Φm is generated by using the mean value μ of output probabilities and a corresponding dispersion σ squared. In the equations, m indicates the cluster number, and N indicates the total number of initial clusters belonging to the class Φm.

$$\mu_m = \Big(\sum_{k \in \Phi m} \mu_k\Big)/N \qquad (1)$$

$$\sigma_m^2 = \Big(\sum_{k \in \Phi m} \sigma_k^2 + \sum_{k \in \Phi m} \mu_k^2 - N \cdot \mu_m^2\Big)/N \qquad (2)$$

Next, obtained at Step S5 are an initial cluster φp, among the initial clusters φi belonging to the new cluster Φm, remotest from the cluster Φm, and an initial cluster φq remotest from the initial cluster φp. As the distance scale d (φp, φq) between the two initial clusters, a Kullback information quantity, a Chernoff distance, a normalized Euclid distance, a Euclid distance, or the like may be used. In this embodiment, a Bhattacharyya distance is used which can be calculated by the following equation (3) in the case of a single Gaussian distribution.

$$d(\phi_p,\phi_q)= \frac{1}{8}(\mu_p-\mu_q)^t\Big(\frac{\Sigma_p+\Sigma_q}{2}\Big)^{-1}(\mu_p-\mu_q) + \frac{1}{2}\ln\frac{|(\Sigma_p+\Sigma_q)/2|}{|\Sigma_p|^{1/2}|\Sigma_q|^{1/2}} \qquad (3)$$

where μi and Σi indicate a mean value and a dispersion, respectively.

Next, at Step S6, the initial clusters φi belonging to the cluster Φm are divided into new clusters Φm and Φ(m+1) nearer to the initial clusters φp and φq obtained at Step S5.

The above process will be described with reference to Fig. 4. In an acoustic space 401, assuming that the cluster Φm is positioned generally at the center of the acoustic space 401 and the cluster φp is positioned near at the right end of the acoustic space 401, then the cluster φq is positioned near at the left end of the acoustic space 401. If the initial clusters φi are divided into the new two clusters nearer to the initial clusters φp and φq, the acoustic space 401 is divided at generally the center thereof into two spaces and the total number M of new clusters is two.

At Step S7, K-means clustering is performed for the new clusters Φi by using all the initial clusters. This K-means clustering is performed until a preset number of iterations is performed or the total distortion Dm becomes a threshold value or smaller, to search a cluster Φd having a maximum total distortion, and d is set to m to return to Step S3.

The total distortion of each cluster can be obtained by the following equation (4).

$$D_m = \sum_{i \in \Phi m} d(\Phi_m, \phi_i) \qquad (4)$$

If the total number M of clusters exceeds the preset number (e.g., 600), the process is terminated. In this manner, the shared state of M clusters can be determined.

BNSDOCID: <EP__ 0750293A2 I_>

(3) Determination of a state shared structure of Triphone HMMs (103)

Each state of the triphone HMMs designed at Design of Initial Clusters (101) is assigned a nearest cluster among the clusters designed at Top-down Clustering (102) to determine the state shared structure of triphone HMMs by using the shared state numbers. For judgement of a distance, the Bhattacharyya distance was used and the states were assigned. In this manner, the acoustically nearer states are shared between triphone HMMs or in a single triphone HMM.

In Fig. 4, a symbol such as /a·Z·i/ indicates a single triphone. In the example shown in Fig. 4, a model having three states is shown. This triphone is a phone "Z" having a right phone "1" and a left phone "a". For example, in Fig. 4, the first states of /a·Z·i/, /a·Z·y/, and /a·Z·a/ are represented by the same state 402, the second states of /a·Z·i/ and /a·Z·y/ are represented by the same state 403, and only the second state of /a·Z·a/ is represented by another state 404. All the first to third states of /a·Z·i/ and /a·Z·y/ are shared by the same state, and so they cannot be discriminated. However, for example, the phone series and triphones of "azia" and "azya" are as follows.

azia (phones)	a	z	i	a
	qAz	aZi	zIa	iAq
azya (phones)	a	z	y	a
	qAz	aZy	zYa	yAq

A silent portion without a phone is represented by q. Since qAz, aZi, and aZy have the same shared state, the two words "azia" and "azya" cannot be discriminated at this point. However, if zIa and zYa, or iAq and yAq, have not the same state shared structure, the two words can be discriminated at one of these points and there is no problem of practical recognition processes.

In some case (particularly if the total number of shared states is small), all states of triphones having the different middle phones may share the same state. In such a case, if division is necessary, all triphones can be modified to have different acoustic characteristics by assigning a shared state number obtained by adding 1 to the total shared state number, to the state (e.g., middle state) of each triphone to become discriminable.

(4) Learning state shared triphone HMMs (104)

In accordance with the state shared structure determined at (3), the states of triphones are tied to one for performing tied-state learning. This learning may use conventional methods such as EM-algorithm.

Fig. 5 is a block diagram illustrating a voice recognition process used by the voice recognition apparatus of the invention.

In this embodiment, HMMs 505 are generated by the above described procedure 510. A voice section is extracted by an extractor 501 from a voice signal input from a microphone or the like. The extracted voice signal is analyzed by an acoustic analyzer 502. A likelihood calculator 503 obtains a likelihood of each state of HMMs 505. By using the obtained likelihood, a grammar 506, and a voice recognition network 507, a language searcher 504 searches a language series having the largest likelihood and outputs it as the voice recognition results.

Fig. 6 shows the results of recognition of 100 sentences spoken by 10 arbitrary speakers, the recognition being made by using grammars constituted by 1000 words and the voice recognition apparatus of the embodiment. In Fig. 6, a sentence recognition rate (%) indicates a percentage of sentences whose input voices were all correctly recognized, and a word recognition rate (%) is a percentage of correctly recognized words in a spoken sentence.

As above, with the voice recognition performed by using the state shared structure with 600 shared states in total generated by the procedure of the first embodiment, sentence and word recognition rates much higher than conventional phone HMM, right-context HMM, and triphone HMM were obtained.

Next, the second embodiment of the invention will be described.

The above-described clustering algorithm uses a distance scale considering the dispersion σ. Therefore, if the number of initial clusters ϕi and the number of final clusters are very large, the calculation amount is immense. Therefore, if a distance calculation requiring a large calculation amount for calculating the distances between all clusters is used, a correspondingly longer time is required. In view of this, two calculation types, a simple distance calculation and an accurate distance calculation for calculating accurate distances, are used. The simple distance calculation is used for clusters of a first group starting from the first cluster to an intermediate cluster among the total number of clusters, whereas the accurate distance calculation is used for clusters including the cluster next to the intermediate cluster to the final cluster. In this manner, the time required for distance calculation is shortened and the process can be speeded up. In this second embodiment, the simple distance calculation uses the Euclid distance and the accurate distance calculation uses the Bhattacharyya distance.

Fig. 7 is a flow chart illustrating processes by the second embodiment.

INSDOCID· <EP 0750293A2 I >

First, at Step 701, a cluster Φm containing all initial clusters φi is generated. This corresponds to Step S2 in Fig. 3. At Step 701 it is checked whether the total number M of clusters have been obtained. If smaller than M, the procedure continues, and if M, the procedure is terminated. At Step 703, it is judged whether the next clustering uses the simple distance calculation or the accurate distance calculation. If the number (m) of clusters is smaller than total number M (e.g., 600) of clusters subtracted by x (e.g., 10), i.e., from the first cluster to the 590-th cluster, the flow advances to Step 704 to execute clustering with the simple distance calculation.

If the number (m) of clusters is (M - x) or larger, the flow advances to Step 705 to execute clustering with the accurate distance calculation to the final cluster M. The processes at Steps 704 and 705 are different in there calculation methods and correspond to Steps S4 to S7 of Fig. 3. Namely, Step 705 uses the Bhattacharyya distance and are the same processes at Steps S4 to S7 of Fig. 3, and Step 704 uses the Euclid distance and calculates the distances at Steps S4 to S7 by the Euclid distance. After Step 704 or 705, one cluster is added at Step 706 and the flow returns to Step 702.

The distance calculation in this embodiment may use other distances different from the Bhattacharyya distance and Euclid distance.

In the above embodiments, HMM is used as the voice recognition model. Instead of HMM, other models may be used if they are state transition models having distributions. Although the triphone is used as a model unit, the recognition unit may be music or other information.

In the above embodiments, although voice recognition is used, the above embodiment procedures are applicable to model design of pattern recognition by using models having similar distributions.

The invention is applicable to a system having a plurality of equipments and to a single equipment. The invention is applicable to a program embodying the invention and supplied to a system or equipment.

As described so far, the features of the embodiments reside in (1) that clusters are generated through top-down clustering considering the whole acoustic space, (2) that states can be shared between phone classes and in each phone class, and (3) a state shared structure of an arbitrary speaker can be generated directly. Therefore, a triphone HMM of an efficient state shared structure can be designed through top-down clustering. By using the voice recognition model designed by the procedures of the invention, high speed and high performance voice recognition can be realized.

Claims

1. A method of processing signals representative of speech samples to generate a state transition model in which a state shared structure of the state transition model is determined, the method comprising:

    a step of setting the states of a triphone state transition model in an acoustic space as initial clusters;
    a clustering step of generating a cluster containing said initial clusters by top-down clustering;
    a step of determining a state shared structure by assigning a short distance cluster among clusters generated by said clustering step, to the state transition model; and
    a step of learning a state shared model by analyzing the states of the triphones in accordance with the determined state shared structure.

2. A method according to claim 1, wherein said clustering step executes clustering to generate a predetermined number of clusters by simple distance calculation, and after generating the predetermined number of clusters, to generate clusters by accurate distance calculation.

3. A method according to claim 2, wherein said accurate distance calculation uses a Bhattacharyya distance.

4. A method according to claim 2, wherein said accurate distance calculation uses a Euclid distance.

5. A method according to claim 1, 2, 3 or 4, wherein said clustering step is defined by an output probability of states.

6. A voice recognition apparatus using a state transition model, comprising:

    input means for inputting voice information;
    analyzing means for analyzing the voice information input from said input means;
    likelihood calculation means for calculating a likelihood between the voice information analyzed by said analyzing means and the state transition model; and
    output means for outputting as a recognition result a language series having a largest likelihood determined by said likelihood calculation means,

BNSDOCID: <EP 0750293A2 I >

wherein the state transition model is a model obtained by: setting the states of a triphone state transition model in an acoustic space as initial clusters; generating a cluster containing said initial clusters by top-down clustering; determining a state shared structure by assigning a short distance cluster among clusters generated by said clustering step, to the state transition model; and learning a state shared model by analyzing the states of the triphones in accordance with the determined state shared structure.

7. A voice recognition apparatus according to claim 6, wherein the top-down clustering generates a recognition model by executing clustering to generate a predetermined number of clusters by simple distance calculation, and after generating the predetermined number of clusters, to generate clusters by accurate distance calculation.

8. A voice recognition apparatus according to claim 7, wherein said accurate distance calculation uses a Bhattacharyya distance.

9. A voice recognition apparatus according to claim 7, wherein said accurate distance calculation uses a Euclid distance.

10. A voice recognition apparatus according to claim 6, 7, 8 or 9, wherein the top-down clustering is defined by an output probability of states.

11. A voice recognition method using a state transition model, comprising:

an input step of inputting voice information;
an analyzing step of analyzing the voice information input from said input means;
a likelihood calculation step of calculating a likelihood between the voice information analyzed by said analyzing means and the state transition model; and
an output step of outputting as a recognition result a language series having a largest likelihood determined by said likelihood calculation means,
wherein the state transition model is a model obtained by: setting the states of a triphone state transition model in an acoustic space as initial clusters; generating a cluster containing said initial clusters by top-down clustering; determining a state shared structure by assigning a short distance cluster among clusters generated by said clustering step, to the state transition model; and learning a state shared model by analyzing the states of the triphones in accordance with the determined state shared structure.

12. A method of generating a state transition model in which a state shared structure of the state transition model is determined, the method comprising:

the step of arranging the states of a transition model in an acoustic space as an initial cluster;
the step of iteratively dividing said states in said acoustic space into a number of sub-clusters; and
the step of determining a state shared structure by grouping acoustically similar clusters and assigning them to the state transition model.

13. A method according to claim 12, wherein the transition model is a triphone state transition model.

14. A method according to claim 12 or 13, further comprising the step of learning a state shared model by analysing the states of the transition model in accordance with the determined state shared structure.

15. A method of generating a state transition model in which a state shared structure of the state transition model is determined, the method comprising:

a step of setting the states of a triphone state transition model in an acoustic space as initial clusters;
a clustering step of generating a cluster containing said initial clusters by top-down clustering;
a step of determining a state shared structure by assigning a short distance cluster among clusters generated by said clustering step, to the state transition model; and
a step of learning a state shared model by analysing the states of the triphones in accordance with the determined state shared structure.

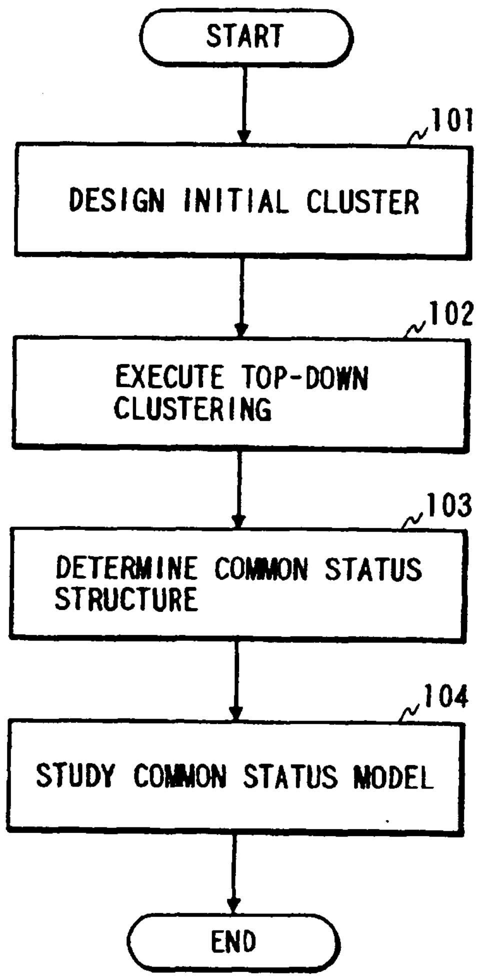
16. A data carrier programmed with instructions for carrying out the method according to any of claims 1 to 5 or 11 to 15.

BNSDOCID: <EP 0750293A2_I_>

17. A data carrier conveying a state transition model as generated by a method according to any one of claims 1 to 5 or 12 to 15.

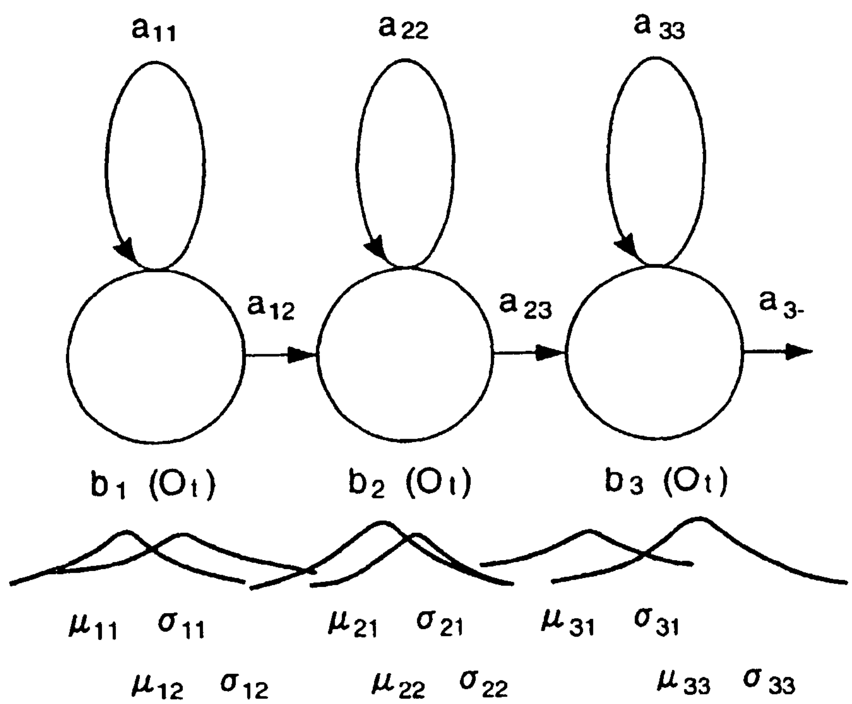
BNSDOCID <EP _0750293A2_I >

FIG. 1



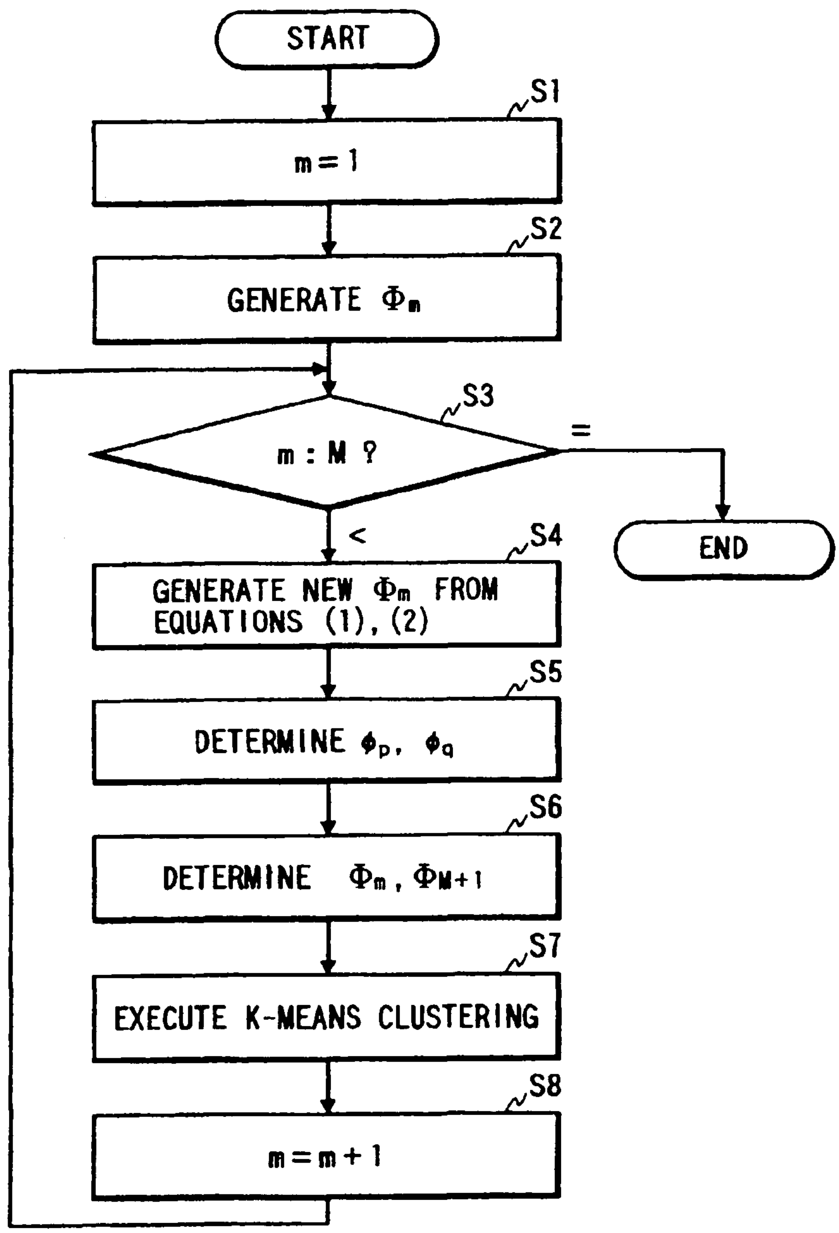
3NSDOCID: <EP___0750293A2_I_>

FIG. 2



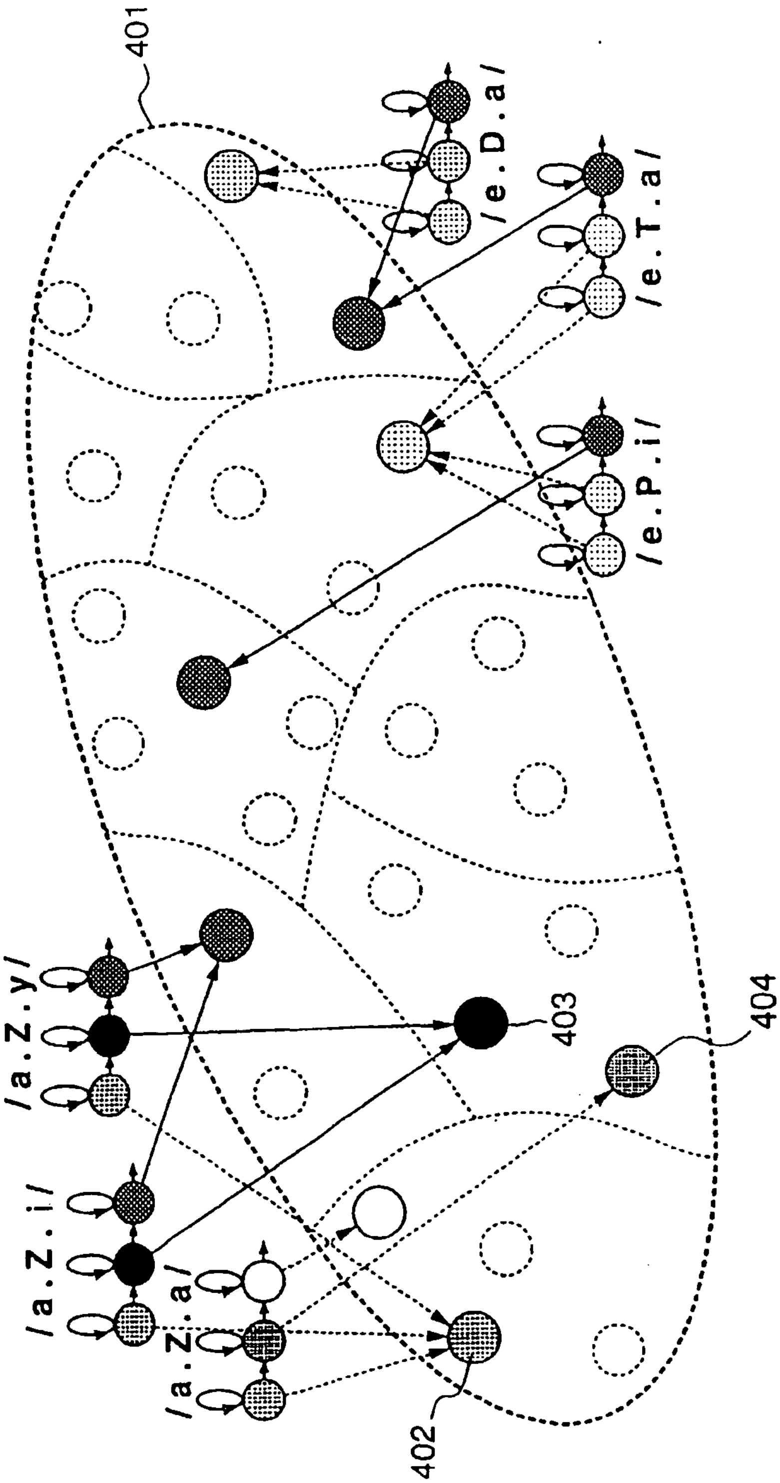
BNSDOCID: <EP 0750293A2 I >

FIG. 3



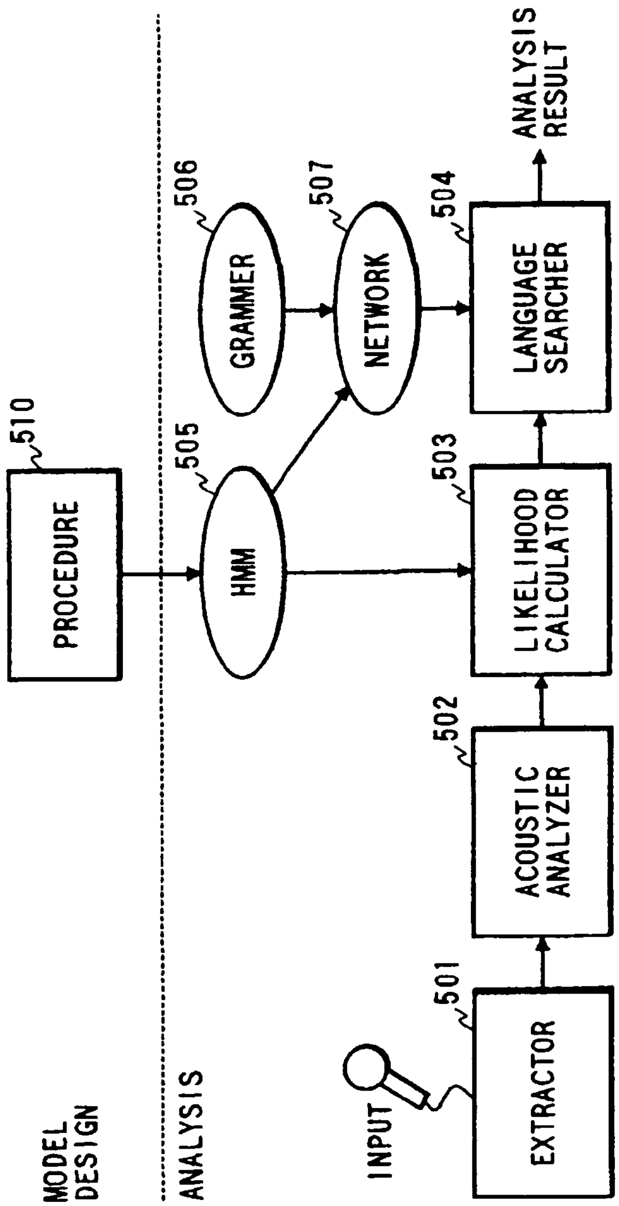
BNSDOCID: <EP___0750293A2_I_>

FIG. 4



BNSDOCID: <EP 0750293A2 I >

FIG. 5



INSDOCID: <EP__ 0750293A2_I_>

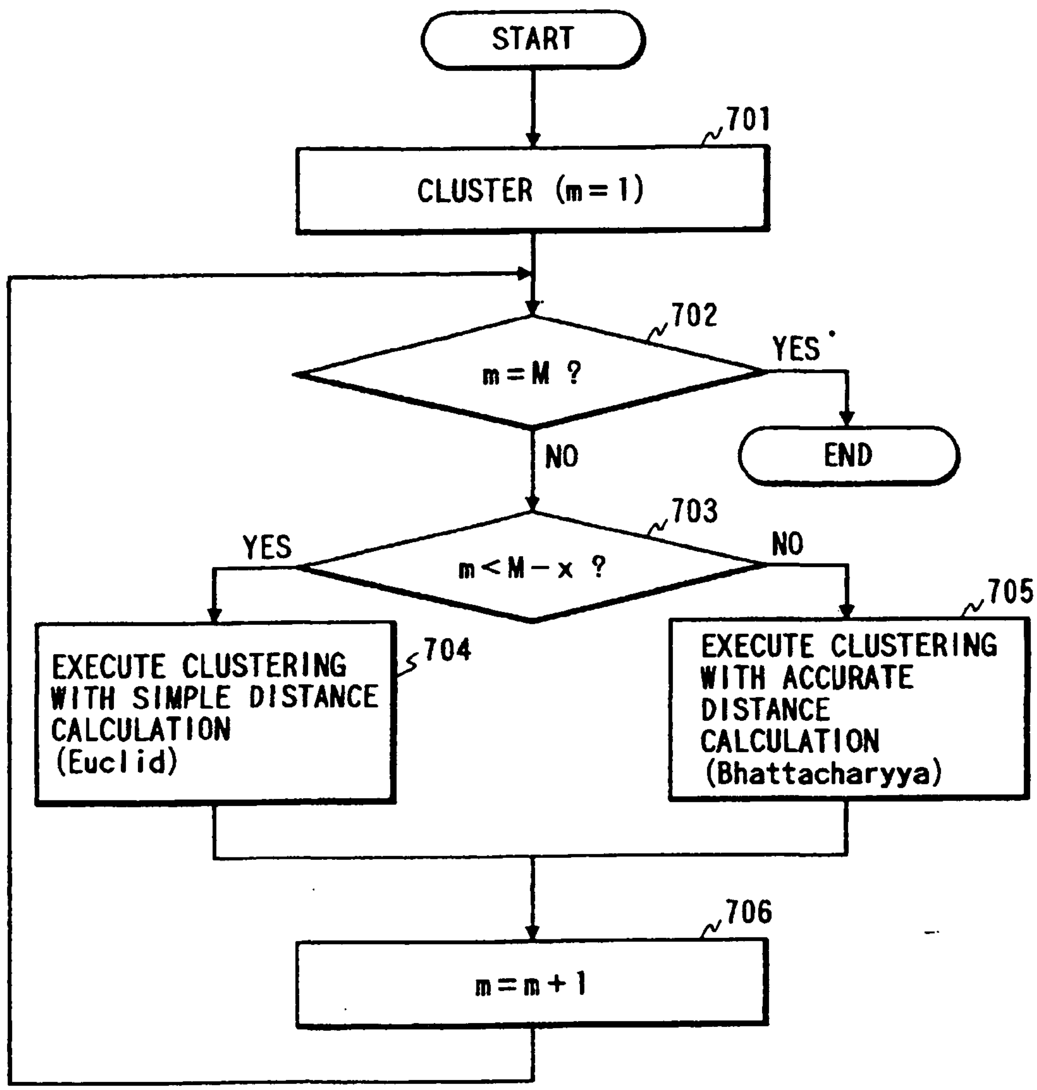
FIG. 6

RATE OF SUCCESSFUL ANALYSIS (%)

HMN	NO. OF STATUS	SENTENCE (%)	WORD (%)
PHONE	75	76.4	93.9
RIGHT-CONTEXT	711	85.0	96.2
TRIPHONE	9927	86.2	96.9
NO. OF COMMON STATUS 600	600	86.6	97.3

BNSDOCID <EP 0750293A2 I >

FIG. 7



NSDOCID: <EP__ 0750293A2_I_>

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 750 293 A3

(12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
08.10.1997 Bulletin 1997/41

(43) Date of publication A2:
27.12.1996 Bulletin 1996/52

(21) Application number: 96304526.5

(22) Date of filing: 18.06.1996

(51) Int Cl.6: G10L 5/06, G10L 7/08, G10L 9/06

(84) Designated Contracting States:
DE FR GB IT

(30) Priority: 19.06.1995 JP 151489/95

(71) Applicant: CANON KABUSHIKI KAISHA
Tokyo (JP)

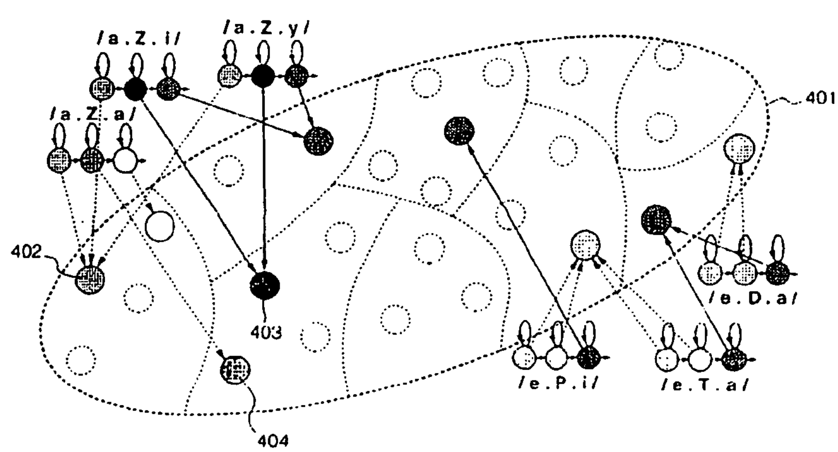
(72) Inventors:
- Komori, Yasuhiro
Ohta-ku, Tokyo (JP)
- Ohora, Yasunori
Ohta-ku, Tokyo (JP)

(74) Representative:
Beresford, Keith Denis Lewis et al
BERESFORD & Co.
2-5 Warwick Court
High Holborn
London WC1R 5DJ (GB)

(54) **State transition model design method and voice recognition method and apparatus using same**

(57) An object of the invention is to provide a method of generating a state transition model capable of high speed voice recognition and to provide a voice recognition method and apparatus using the state transition model. To this end, a method is provided which generates a state transition model in which a state shared structure of the state transition model is designed, the method including a step of setting the states of a triphone state transition model in an acoustic space as initial clusters, a clustering step of generating a cluster containing the initial clusters by top-down clustering, a step of determining a state shared structure by assigning a short distance cluster among clusters generated by the clustering step, to the state transition model and a step of learning a state shared model by analyzing the states of the triphones in accordance with the determined state shared structure.

FIG. 4



EP 0 750 293 A3

INSDOCID: <EP__0750293A3_I_>

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 96304526.5

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl. 6)
A	EP - A - 0 362 785 (NEC CORPORATION) * Fig. 1,2,3; abstract; claim 1; column 6, lines 21-28 * --	1-17	G 10 L 5/06 G 10 L 7/08 G 10 L 9/06
A	US - A - 5 165 007 (BAHL et al.) * Abstract; fig. 1-4; claims 1,2; column 6, line 29 - column 7, line 47 * --	1-17	
A	EP - A - 0 237 934 (TOSHIBA K.K.) * Abstract; fig. 1-5; claim 1 * ----	1-17	
			TECHNICAL FIELDS SEARCHED (Int. Cl.6)
			G 10 L 5/00 G 10 L 7/00 G 10 L 9/00 G 10 L 3/00

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
VIENNA	31-07-1997	BERGER

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P0401)

BNSDOCID: <EP 0750293A3 I >

(19) Europäisches Patentamt
European Patent Office
Offi e uropéen des brevets

(11) EP 1 204 091 A3

(12) **EUROPEAN PATENT APPLICATION**

(88) Date of publication A3:
19.11.2003 Bulletin 2003/47

(51) Int Cl.7: **G10L 15/06**

RECEIVED
08 ... 2003
MARK... CLERK

(43) Date of publication A2:
08.05.2002 Bulletin 2002/19

(21) Application number: **01309333.1**

(22) Date of filing: **02.11.2001**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **02.11.2000 US 245139 P**
01.11.2001 US 998959

(71) Applicant: **AT&T Corp.**
New York, NY 10013-2412 (US)

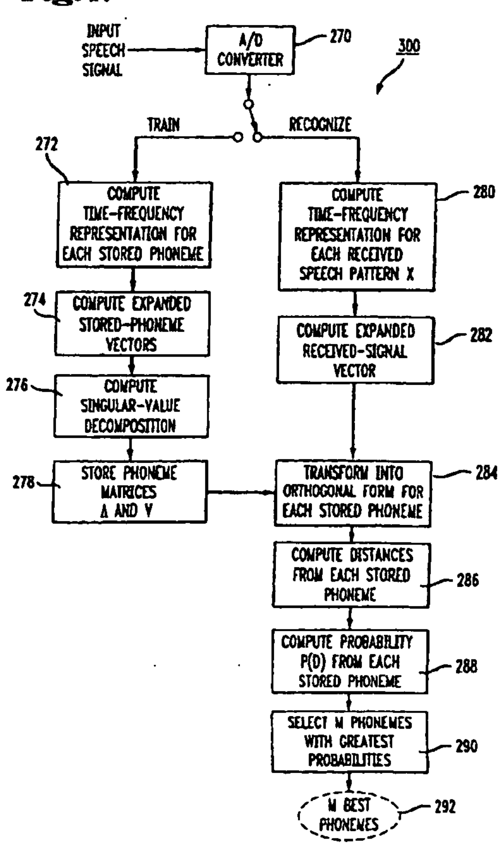
(72) Inventor: **Atal, Bishnu Saroop**
New Providence, New Jersey (US)

(74) Representative: **Harding, Richard Patrick**
Marks & Clerk,
4220 Nash Court
Oxford Business Park South
Oxford OX4 2RU (GB)

(54) **A system and method of pattern recognition in very high-dimensional space**

(57) A system and method of recognizing speech comprises an audio receiving element and a computer server. The audio receiving element and the computer server perform the process steps of the method. The method involves training a stored set of phonemes by converting them into n-dimensional space, where n is a relatively large number. Once the stored phonemes are converted, they are transformed using single value decomposition to conform the data generally into a hypersphere. The received phonemes from the audio-receiving element are also converted into n-dimensional space and transformed using single value decomposition to conform the data into a hypersphere. The method compares the transformed received phoneme to each transformed stored phoneme by comparing a first distance from a center of the hypersphere to a point associated with the transformed received phoneme and a second distance from the center of the hypersphere to a point associated with the respective transformed stored phoneme.

Fig. 14



EP 1 204 091 A3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 01 30 9333

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	US 5 638 489 A (TSUBOKA EIICHI) 10 June 1997 (1997-06-10) * column 1, line 10-51 * * column 7, line 11-32; figure 3 *	5,14,17, 18	G10L15/06
A	LILLY B T ET AL: "Robust speech recognition using singular value decomposition based speech enhancement" TENCON '97. IEEE REGION 10 ANNUAL CONFERENCE. SPEECH AND IMAGE TECHNOLOGIES FOR COMPUTING AND TELECOMMUNICATIONS., PROCEEDINGS OF IEEE BRISBANE, QLD., AUSTRALIA 2-4 DEC. 1997, NEW YORK, NY, USA,IEEE, US, 2 December 1997 (1997-12-02), pages 257-260, XP010264234 ISBN: 0-7803-4365-4 * abstract * * page 259, left-hand column, paragraph 3.2 * * page 260, left-hand column, paragraph 5.0 *	1-6, 17-19	
A	EP 0 750 293 A (CANON KK) 27 December 1996 (1996-12-27) * page 2, line 3-5 * * page 3, line 21-36 * * page 5, line 33-39; figures 1,5 *	5,14,17, 18	TECHNICAL FIELDS SEARCHED (Int.Cl.7) G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
MUNICH	29 September 2003	Greiser, N

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons
& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 01 30 9333

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

29-09-2003

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5638489 A	10-06-1997	JP 2795058 B2	10-09-1998
		JP 5333898 A	17-12-1993
		US 5608841 A	04-03-1997
		US 5608840 A	04-03-1997
EP 0750293 A	27-12-1996	JP 9006386 A	10-01-1997
		EP 0750293 A2	27-12-1996
		US 5812975 A	22-09-1998

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82